Valeur et Véracité de la donnée : enjeux pour l'entreprise et défis pour le Data Scientist.

Bruno Teboul, Thierry Berthier

Valeur et Véracité de la donnée : Enjeux pour l'entreprise et défis pour le Data Scientist

Bruno Teboul,
Directeur scientifique, R&D et Innovation du groupe Keyrus,
Membre de la Gouvernance de la Chaire Data Scientist de l'École Polytechnique
Doctorant et enseignant à l'Université Paris-Dauphine

Thierry Berthier
Maître de conférences en mathématiques, Université de Limoges,
Chaire de cybersécurité & cyberdéfense Saint-Cyr Thales

Actes du colloque « La donnée n'est pas donnée »
École Militaire – 23 mars 2015

Les mégadonnées ou big data ont investi l'ensemble des activités industrielles, économiques et sociales. Leurs exploitations modifient en profondeur notre rapport au monde et fournissent des clés fondamentales qui permettent de repousser l'incertitude, de comprendre et parfois de contrôler des phénomènes complexes. En transformant les espaces numériques et physiques, le déluge des données influence et oriente les pratiques économiques, sociales et politiques. De nouvelles échelles temporelles et spatiales deviennent ainsi accessibles à l'interprétation et la compréhension humaine tout en apportant toujours plus de défis technologiques. Qu'elle soit créée par l'individu ou par l'infrastructure connectée, la donnée fournit aujourd'hui une image de plus en plus précise du réel. L'enjeu majeur réside alors dans l'exploitation de cette donnée et dans notre capacité à tirer partie d'un ensemble de données massives pour améliorer un processus, construire une prévision de réalisation d'un événement, établir une recommandation de produits, de services (prescription), ou encore optimiser l'aide à la décision. Les technologies du big data ne se contentent pas de faire parler la donnée, elles créent du sens et produisent des solutions performantes [1]. Cette puissance d'analyse des mégadonnées s'appuie naturellement sur un prérequis de qualité : la véracité et la valeur d'une donnée conditionnent l'ensemble du processus d'exploitation. Elles en sont le fondement et doivent faire l'objet d'un questionnement systématique.

1- La qualité des données

Gartner et IBM utilisent six lettres V fondamentales pour décrire le big data : Volume, Variété, Vélocité, Visibilité, Valeur et Véracité.
Le volume est lié aux multiples sources de production des données. Qu'il s'agisse de données d'entreprises, de données publiques, de données issues de transactions, de données produites par des capteurs automatisés, des objets connectés ou publiées sur les médias sociaux, ces informations sont toujours collectées et stockées sur des supports numériques sous forme de fichiers binaires. Leur volume est donc facilement calculable. Ainsi, la production mondiale de données atteindra en 2020 les 40 Zo (un Zo est égal à dix puissance vingt et un octets) [2]. L'évolution de cette production est exponentielle : 90% des données actuelles ont été produites durant les deux dernières années. L'émergence de villes connectées puis de villes ubiquitaires [3] renforce cette tendance. En installant l'information ubiquitaire, une information accessible à tous, partout, en permanence, les objets connectés et la géolocalisation contribuent massivement au déluge de données.

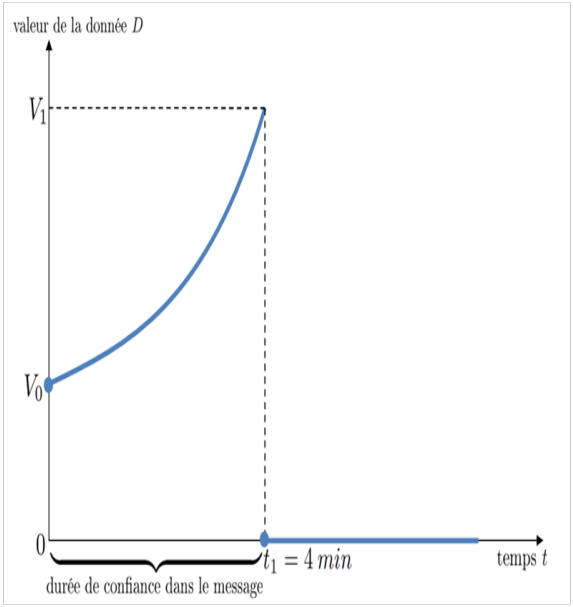
La variété résulte des sources de données hétérogènes, souvent non ou peu structurées (données de capteurs, données de géolocalisation, sons, vidéos, textes,...). Cette variété a motivé la construction de systèmes capables de « gérer » la non structuration (NoSql, Hadoop,...) tout en assurant une meilleure répartition de la charge des volumes sur l'infrastructure de calcul. Comme le volume, la variété se mesure simplement en dénombrant les différents formats présents dans le corpus big data à traiter. Les projections algorithmiques individuelles [4] participent à la variété de la donnée.
La vélocité intervient dans les contextes de données en mouvement, de « data streaming » et de traitement temps réel de ces données. Elle est liée à la vitesse de production de la source, au flux, au débit et à la vitesse de collecte du système. Ici encore, la vélocité est une grandeur facilement mesurable.
La **visibilité** des données dépend fortement du support de stockage, du caractère ouvert ou non de l'information et de l'efficacité des algorithmes de collecte et autres crawler [5]. On pourrait compléter ces quatre premiers V par celui de la variabilité de la donnée dans certains contextes. Cette variabilité s'exprime pour des données dont le contenu évolue dans le temps et l'espace. Ces évolutions produisent alors de nouvelles données indicées par le temps.

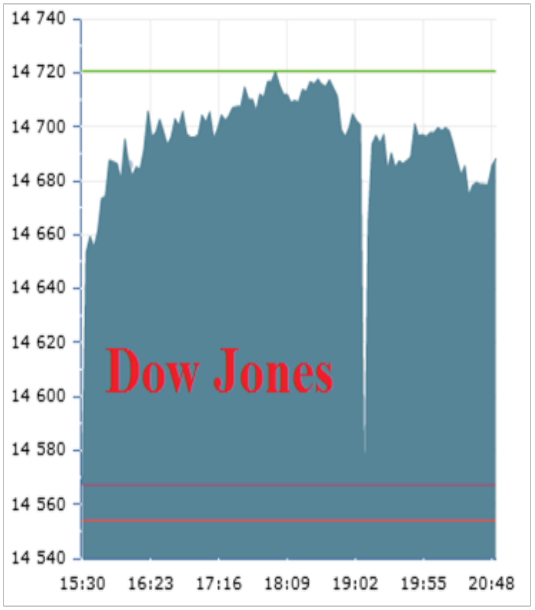
Les deux derniers V désignent la **valeur** et la **véracité** d'une donnée. Ces qualités sont beaucoup plus complexes à définir et à mesurer que les quatre premières [6]. La **valeur** recouvre en effet plusieurs spectres nécessitant chacun une analyse spécifique. On parlera ainsi de valeur d'impact sur un contexte, de valeur de modélisation, de valeur de prédiction, de valeur de management, de valeur économique ou de revente... La **véracité** conditionne quant à elle directement la pertinence de la donnée. Si des données incertaines peuvent être traitées au même titre que des données « certifiées », leur interprétation dans le cadre de fausses données peut engendrer de fortes turbulences sur l'ensemble des systèmes associés et provoquer des sinistres conséquents lorsque des décisions sont prises sur la base de cette interprétation. En fait, il n'existe pas de valeur « absolue » d'une donnée mais plutôt des valeurs relatives à un contexte d'interprétation, à un instant donné. La valeur d'impact d'une donnée peut ainsi être totalement indépendante de sa véracité sur le contexte comme ont pu le montrer des opérations de hacking capables de déstabiliser les grands indices boursiers sur la seule base de fausses données publiées.

2- Valeur instantanée d' interprétation d'une donnée.

Fixons quelques notations utiles : une donnée numérique D est une suite binaire finie, c'est-à-dire un mot de longueur finie formé de 0 et de 1. Un programme P prenant D en entrée calcule la sortie P<D>. Ce programme est exécuté sur un système de calcul S [1]. Un contexte C réunit des acteurs humains et des systèmes de calculs autour d'un ensemble d'objectifs économiques ou stratégiques [7].

La valeur d'interprétation [8] d'une donnée D par un programme P relativement au contexte C est une fonction qui à l'instant t associe $Val_t(D/P, C)$. Cette valeur dépend de l'instant d'évaluation t, du programme P prenant D en entrée et calculant P<D> et du contexte d'évaluation C.
Si le programme P ne fait que réécrire la donnée sans la modifier alors P<D> = D, et
$Val_t(D/P, C) = Val_t(D/C)$ qui désigne dans ce cas la valeur de la donnée brute n'ayant subi aucun traitement.
Si au contraire P « raffine » D en supprimant par exemple le bruit contenu, en supprimant les doublons ou en donnant à D un format pertinent pour le contexte C alors l'exécution du programme P apporte de la valeur par rapport à la donnée brute et $Val_t(D/P, C) > Val_t(D/C)$.
Ce sont toujours les acteurs et les systèmes du contexte C qui évaluent et fixent la valeur d'interprétation de la donnée. Lorsque la véracité de la donnée est établie, sa valeur d'interprétation augmente. Si P est un programme certifiant la véracité de la donnée, alors, lorsque P confirme la véracité de D à l'instant t, on a encore $Val_t(D/P, C) > Val_t(D/C)$.
En revanche, lorsque P ne parvient pas à certifier cette véracité, $Val_t(D/P, C)$ reste inférieure ou égale à la valeur brute $Val_t(D/C)$. La donnée peut être fausse ou vraie sans que le programme P ne soit en mesure de le détecter. La valeur instantanée d'interprétation peut alors présenter de forts gradients comme le montre l'exemple emblématique du faux tweet d'Associated Press et son impact sur les indices boursiers en 2013. Le 20 avril 2013, une cellule de hackers agissant sous le nom d'Armée Électronique Syrienne (SEA) [9] attaque le compte officiel Twitter de l'agence Associated Press. Elle en prend momentanément le contrôle et publie à 13h07 le message suivant : « Breaking : Two Explosions in the White House and Barack Obama is injured ». Les 1,9 millions d'abonnés au compte twitter d'Associated Press découvrent le faux message posté par la SEA en le considérant comme authentique (il n'y a en effet aucune raison de douter de sa véracité à cet instant) . La réaction des marchés est presque immédiate. Entre 13h08 et 13h10, l'indice principal de Wallstreet, le Dow Jones (DJIA) perd immédiatement 145 points soit l'équivalent de 136 milliards de dollars. Le mouvement de panique touche autant les opérateurs humains que les systèmes de trading algorithmique (HFT) qui interprètent et réagissent au faux tweet. Les actions Microsoft, Apple, Mobil perdent plus de 1% de leur valeur presque instantanément. Quelques minutes plus tard, l'agence Associated Press reprend le contrôle de son compte officiel et publie un démenti annonçant que le message précédent était faux et qu'il résultait d'un piratage informatique. Les indices retrouvent alors rapidement leurs valeurs initiales. Cet exemple montre que la publication d'une donnée fausse longue de soixante caractères suffit à provoquer un « Flash crash » et un mouvement de 136 milliards de dollars sur les marchés mondiaux...





La courbe ci-dessus décrit la valeur d'interprétation instantanée de la donnée produite par la SEA durant les quatre premières minutes suivant la publication du faux tweet et montre la discontinuité de cette valeur une fois le démenti publié. La valeur d'interprétation croit exponentiellement d'une valeur initiale V0 jusqu'à un maximum V1 puis s'annule lorsque l'agence AP reprend le contrôle de son compte et révèle l'agression. La donnée brute issue de la cyberattaque possède alors une forte valeur d'interprétation durant quatre minutes et une valeur d'impact encore plus importante puisqu'elle s'élève à plus de 136 milliards de dollars sur le contexte des marchés boursiers. On constate que la valeur d'impact d'une donnée peut être indépendante de sa véracité. C'est avant tout le crédit ou la confiance que l'on accorde à une donnée qui permet de fonder sa valeur d'impact. L'exemple du faux tweet de la SEA illustre la puissance et la complexité des interactions qui existent et qui opèrent entre la valeur et la véracité de la donnée. Il montre alors tout l'intérêt, pour le contexte d'interprétation, à privilégier des données certifiées lorsque la véracité de ces données est « calculable » durant la phase de collecte.

Si la valeur instantanée d'interprétation offre bien une approche individuelle de la valeur d'une donnée, le cadre des mégadonnées requiert quant à lui une prise en compte globale. L'ensemble massif de données possède une valeur fondée sur les relations internes liant ces données, sur leur sens collectif et sur l'information résultant de cet ensemble. Le gain obtenu après exploitation du jeu de données lui confère sa valeur globale [10],[11].

3- Valeur d'un ensemble massif de données.

Nous nous plaçons désormais dans un cadre « big data » et allons définir la valeur d'un ensemble massif de données fondée sur le gain obtenu après exploitation de cet ensemble par un système de calcul. Dans toute la suite, on note D un ensemble de données $D = \{ D_1, D_2,....,D_n \}$. Ceci suppose que n est grand, suffisamment grand pour que la suppression d'une des données de l'ensemble n'ait pas d'influence sur le résultat du calcul exploitant cet ensemble.

L'ensemble D est exploitable par un système de calcul S réunissant des algorithmes et des machines. Pour une organisation (entreprise, administration, institution, laboratoire), une ligne de contrainte L désigne un sous-domaine d'activité de l'organisation qui influence directement son fonctionnement, son efficacité et sa rentabilité. La ligne de contrainte peut être temporelle lorsqu'elle se rapporte au temps nécessaire à un processus de production ou spatiale lorsqu'il s'agit d'une distance ou d'une surface à prospecter. L peut aussi concerner un effectif en ressources humaines (le nombre d'ingénieurs affectés à un projet) ou encore le coût de développement d'un objet ou d'un service. On associe à une ligne de contrainte L sa fonction $C_L(t)$ d'évaluation qui dépend de l'instant t de la mesure. La fonction dévaluation peut par exemple désigner le coût d'un processus, sa durée, ou encore un volume, une surface, une longueur ou l'effectif de personnels intervenant dans ce processus.

La gain obtenu sur la ligne de contrainte L après exploitation de l'ensemble de données D par le système de calcul S s'écrit :

$$G_L(D, S) = C_L(\text{ après exploitation de D par S}) - C_L(\text{ avant exploitation de D par S }) = \Delta C_L$$

On est alors en mesure de définir la valeur $V_L(D)$ de l'ensemble de données D sur la ligne de contrainte L en considérant le maximum des gains obtenus lorsque l'on fait varier le système de calcul S qui traite D :

$$V_L(D) = k_L \, \mathbf{max}_S (G_L(D, S))$$

Le facteur k_L est une constante dépendant de la ligne de contrainte L. C'est un coefficient de normalisation défini pour chaque ligne de confiance. Si plusieurs lignes de contraintes sont impactées par le traitement de D, k_L peut aussi représenter le poids que l'on donne à L par rapport aux autres lignes de contraintes. Il permet alors de hiérarchiser les lignes de contraintes.

Notons que calculer la valeur précise de $V_L(D)$ reviendrait à faire tourner tous les systèmes de calcul S sur l'ensemble de données D et à sélectionner celui (ou ceux) qui produisent le meilleur gain sur L. Il s'agit donc d'une définition asymptotique de la valeur d'un ensemble de données avec laquelle on se contente d'une approximation approchant $V_L(D)$ par valeurs inférieures.

Donnons à présent deux cas concrets pour lesquels on approche la valeur d'un ensemble de données dans un contexte de traitement big data.

Cas 1 – Les éoliennes VESTAS
La mise en place d'une analyse big data a permis à la société de développement d'éoliennes Vestas d'optimiser son processus d'identification des meilleurs emplacements pour implanter ses éoliennes [12],[13]. Le traitement big data a engendré une augmentation de la performance de production d'électricité et une réduction des coûts énergétiques associés. Grâce aux mégadonnées météorologiques, Vestas est en mesure de décrire le comportement du vent sur une zone choisie et de fournir une analyse de rentabilisation précise à ses clients. Le système big data Vestas-IBM a induit une réduction de 97 % du temps de réponse sur les prévisions éoliennes passant de plusieurs semaines à seulement quelques heures aujourd'hui. Le coût de production par kilowattheure pour les clients a été réduit ainsi que le coût et l'encombrement informatique associés avec une diminution de plus de 40% de la consommation énergétique. La base de données météorologiques « Vestas-éoliennes » atteint les 24 péta-octets. Le logiciel IBM infoSphere BigInsights fonctionnant sur un système x-iDataPlex a assisté le groupe Vestas dans sa gestion des données météorologiques et de localisation. Le traitement de cet ensemble de mégadonnées a fait diminuer la résolution de base des grilles de données éoliennes passant ainsi d'une aire élémentaire de 27 x 27 kilomètres à une aire de 3x3 kilomètres après calculs. Dans ce traitement de mégadonnées, la ligne de contrainte L correspond à la résolution de base des grilles de données, c'est-à-dire à une surface exprimée en kilomètres carrés. Le gain obtenu à partir de l'exploitation des données réduit l'incertitude spatiale de plus de 90 % et donne un aperçu précis (et presque immédiat) du meilleur site d'implantation de l'éolienne dans la zone étudiée. $G_L(D, S) = -720\ km^2$ soit 98 % de gain de précision après traitement et la valeur du jeu de donnée Vestas-éoliennes vérifie $V_L(D) > 720\ k_L$.

Cas 2 – Le zoo de Cincinnati
Confronté à des difficultés de rentabilité, le zoo américain de Cincinnati (Ohio) [14] s'est orienté vers le traitement big data de ses données clients et des données issues de capteurs déployés au sein des attractions et bâtiments du parc. L'image en temps réel des comportements de la clientèle du zoo a permis d'augmenter de 25 % les dépenses des visiteurs en apportant plus de 350 000 dollars de recettes supplémentaires par an. La compréhension fine des données clients a été appliquée à l'optimisation des ressources humaines du zoo et a libéré du temps pour le personnel désormais disponible sur d'autre postes de rentabilité. Le budget de l'entreprise a ainsi retrouvé son équilibre. Dans ce cas, la ligne de contrainte L correspond aux recettes annuelles en dollars, $G_L(D, S) = 350\ 000$ USD soit 25 % de gain annuel et $V_L(D) > 350\ 000\ k_L$.

4- Véracité de la donnée.

La création de fausses données peut être considérée comme un effet collatéral de l'algorithmisation de l'environnement. L'usage de ces fausses données répond quant à lui à des objectifs variés.

De la fausse donnée pour protéger son anonymat

Selon un rapport traitant de la protection des données privées publié en 2015 par la société de cybersécurité Symantec, 57 % des européens se déclarent inquiets quant à la sécurité de leurs informations personnelles, 81 % estiment que leurs données ont une valeur supérieure à mille euros et 31 % n'hésitent plus à communiquer aux systèmes de fausses données pour protéger leurs données personnelles. Des applications ont été développées pour créer de fausses données dans le but de tromper les applications Android qui sont parfois ressenties comme trop intrusives par l'utilisateur. Xprivacy est un outil qui permet de nourrir les applications Android avec de faux contacts, de fausses coordonnées géographiques, de faux dictionnaires user, de faux presse-papiers, de faux historiques d'appels, de faux SMS. L'objectif affiché par Xprivacy est de créer de fausses données pour mieux protéger sa vie privée. Dans le même esprit, le site FakeNameGenerator permet de construire une base de données sous divers formats (MS SQL, MySQL, IBM DB2, Oracle,...) de 50 000 identités fictives cohérentes incluant nom, âge, nationalité, adresse, profession des profils enregistrés. Le site CloneZone réalise le clonage immédiat de tout site web et propose de modifier ou détourner tout ou partie de son contenu.

De la fausse donnée pour construire une cyberattaque

La production d'ensembles de données fictives cohérents intervient de plus en plus souvent durant la phase d'ingénierie sociale préparant une cyberattaque. Le facteur humain représente en effet la première fragilité exploitée lors d'une agression numérique, bien avant la fragilité des systèmes...

Le principal défi pour l'attaquant consiste alors à créer un espace de confiance entre lui et sa victime puis à tirer partie de cette confiance pour mener à bien l'agression. L' «énergie » à déployer pour installer la confiance augmente régulièrement avec la sensibilisation des usagers aux risques et dangers numériques. Les cyberpièges qui fonctionnent aujourd'hui sont de plus en plus complexes et s'appuient parfois sur des structures de données fictives particulièrement élaborées. Ainsi, la récente opération de cyberespionnage Newscaster [15] a démontré toute la puissance des fausses données pour tromper ses cibles. Newscaster est une cyberopération sophistiquée attribuée à l'Iran qui s'est inscrite dans la durée entre 2012 et 2014 et qui a ciblé plus de 2000 personnes aux États-Unis, en Europe et en Israël. Parmi les victimes de cette agression figurent des officiers supérieurs de l'US Army, des ingénieurs d'industries d'armement, des membres du congrès, des chefs d'entreprises. Newscaster a été à la fois longue, structurée, adaptative et furtive. La première phase de l'opération s'est appuyée sur la construction d'un faux site web d'information intitulé NewsOnLine, implanté sur des serveurs américains sous contrôle de l'attaquant et supervisé par une rédaction d'agence de presse totalement fictive. Un noyau d'une vingtaine de profils fictifs de journalistes américains affectés à la rédaction du site a été déployé sur l'ensemble des grands réseaux sociaux (Facebook, Twitter, LinkedIn). Cette rédaction virtuelle et fictive a ensuite noué des contacts privilégiés avec ses lecteurs puis a prospecté en direction de ses futures cibles pour leur proposer de participer à la rédaction d'articles sur le site. Au fil des mois et des échanges, la confiance s'est installée entre les journalistes fictifs et les contributeurs ciblés. Lorsqu'une cible envoyait un article à la rédaction de NewsOnLine pour publication sur le site, l'échange de fichiers était utilisé par les attaquants pour injecter des spywares (logiciels destinés à collecter de manière furtive les données présentes sur un ordinateur) sur les machines des cibles. Durant plus d'un an, des données sensibles ou classifiées ont été collectées et exfiltrées par les superviseurs de Newscaster, dans la plus stricte discrétion, jusqu'à ce que la présence des spywares finisse par être détectée par les systèmes d'antivirus. Cet exemple emblématique prouve que la confiance *a priori* portée à un ensemble attractif de données fictives peut induire la vulnérabilité, y compris chez des victimes très informées des risques numériques. Cette vulnérabilité de confiance accordée trop facilement s'inscrit pleinement dans le facteur humain et son cortège de biais cognitifs qui « poussent à la faute ». Comment alors s'en prémunir ?

La confiance en une donnée

Qu'il s'agisse de « small data » ou de big data, c'est bien la confiance que l'on accorde à la donnée qui conditionne son usage. Tout comme la valeur, la véracité d'une donnée dépend à la fois de l'instant et du contexte d'évaluation. Dans l'absolu, il faudrait être en mesure de remonter systématiquement à l'émetteur initial (la source) d'une donnée pour s'assurer de son intégrité et de sa conformité. Cette démarche de certification, algorithmiquement coûteuse, ne concerne de fait que le domaine des données sensibles. D'autre part, il ne faut pas exclure une mise en défaut du processus de certification par une opération de hacking non détectée [16]...

Si l'approche pragmatique consiste à évaluer la probabilité de véracité d'une donnée connaissant son émetteur, sa réputation et son historique **p**(D vraie / Émetteur, Réputation, Historique), il faut désormais calculer cette probabilité en tenant compte de l'éventualité d'une cyberattaque sur D, soit : **p**(D vraie / Émetteur, Réputation, Historique, **p**(Hacking(D)) > 0). C'est elle qui exprime la confiance que l'on porte en D.

Conclusion

Le data scientist est désormais confronté à de nouveaux défis à la fois complexes et stratégiques [17]. L'un d'entre eux concerne la détection des corpus de données fictives et la certification des données légitimes. C'est en croisant les compétences de cybersécurité et de sciences des données que l'on fera émerger des expertises de qualiticien de la donnée. Le qualiticien de la donnée s'appuiera sur des architectures algorithmiques capables d'évaluer en temps réel la légitimité d'une donnée et d'alerter lorsque le faux numérique aura été détecté. D'une façon générale, les systèmes doivent évoluer vers plus de résilience face aux cyberattaques. L'antifragilité [18], concept introduit par Nassim Nicholas Taleb en 2013, peut apporter une réponse efficace à la prolifération des données fictives. Dépassant les simples notions de résistance et de résilience, l'antifragilité sous entend une amélioration régulière du système au fil des chocs subis et une capacité à profiter de l'évènement aléatoire pour se renforcer. En matière numérique, l'antifragilité ne peut s'installer qu'à la suite d'une montée en puissance du niveau d'intelligence artificielle embarquée dans le système. Finalement, la qualité de la donnée demeure subordonnée à l'antifragilité du système qui la traite.

Bibliographie

[1] TEBOUL B. , AMRI T. *« Les Machines pour le Big Data : Vers une Informatique Quantique et Cognitive »*, 2014. <hal-01096689v2>

[2] McKinsey Global Institute, « Big Data : The next frontier for innovation, competition, and productivity », May 2011

[3] KEMPF O. et BERTHIER T. - « *Ville connectée et algorithmes prédictifs »*, Actes de la conférence Digital Polis 2015, Paris. (à paraître).

[4] BERTHIER T. - « Projections algorithmiques et cyberespace » R2IE – revue internationale d'intelligence économique – Vol 5-2 2013 pp. 179-195.

[5] TEBOUL B. *« Text Mining, Sentiment Analysis, Big Data »*, 5 avril 2013, Les Echos

[6] Rapport CIGREF *«Enjeux business des données. Comment gérer les données de l'entreprise pour créer de la valeur ? »*, octobre 2014.

[7] TEBOUL B. , PICARD T. *« Uberisation = Économie déchirée ? »*, avril 2015, Éditions Kawa.

[8] BERTHIER T., *Sur la valeur d'une donnée*, Publications de la Chaire de cyberdéfense Saint-*Cyr-Sogeti-Thales – mai 2014*.

[9] KEMPF O. et BERTHIER T. - « *L'armée syrienne électronique : entre cyberagression et guerre de l'information* » RDN – revue de la défense nationale – « Guerre de l'information » Vol. mai 2014.

[10] Report - European Commission, DG CONNECT *« A European strategy on the data value chain »*, 2013

[11] Report - European Commission, High Level Expert Group on Scientific Data, *« Riding the Wave : How Europe can gain from the rising tide of scientific data »*, october 2010,

[12] IBM Report, *« Vestas : Turning climate into capital with big data »*, 2011

[13] IBM Big Data Report, *« A collection of Big Data client success stories »*, pp117, 2012

[14] IBM Report, *« The Case for Business Analytics in Midsize Firms – Cincinnati Zoo »*,pp 7-10, 2012

[15] BERTHIER T., *«Newscaster, l'opération iranienne »*, pp 12-14, Vérification sur Internet : quand les réseaux doutent de tout, novembre 2014, Observatoire géostratégique de l'information, IRIS.

[16] BERTHIER T., *Cyberchronique – Décomposition systémique d'une cyberattaque, dissymétries et antifragilité*, Publications de la chaire de cyberstratégie CASTEX, janvier 2014

[17] TEBOUL B. *« Former des bataillons de Data Scientists à l'Ecole Polytechnique »*, 16 octobre 2014, Silicon, article en ligne, http://www.silicon.fr/bruno-teboul-keyrus-polytechnique-chaire-data-scientist-99428.html

[18] TALEB NN. "*Antifragile : les bienfaits du désordre*" , Editions Les Belles Lettres, 2013